

# Single-cell RNA-Seq data sets

Kushal K Dey, Chiaowen Joyce Hsiao & Matthew Stephens

*Stephens Lab*, The University of Chicago

\*Corresponding Email: [mstephens@uchicago.edu](mailto:mstephens@uchicago.edu)

March 21, 2016

## Abstract

This package includes three single-cell RNA-seq data sets: 1) Mouse spleen cells due to Jaitin et al 2014 [?], 2) Mouse preimplantation embryos due to Deng et al 2014 [?], and 3) Mouse cortex and hippocampus cells data due to Zeisel et al 2015 [?]. Each data set is stored in an *ExpressionSet* objects. We previously analyzed Jaitin et al and Deng et al in the manuscript entitled "Clustering RNA-seq expression data using grade of membership models", Kushal K Dey, Chiaowen Joyce Hsiao, and Matthew Stephens (2016).

**singlecellRNAseqData version:** 0.99.0 <sup>1</sup>

---

<sup>1</sup>This document used the vignette from *Bioconductor* package [DESeq2](#), [cellTree](#), [CountClust](#) as *knitr* template

## Contents

---

### 1 Installation

---

To install the Bioconductor version of this package,

```
source("http://bioconductor.org/biocLite.R")
biocLite("singlecellRNAseqData")
```

To install the working version from Github, the user needs CRAN package [devtools](#).

```
library(devtools)
install_github("kkdey/singlecellRNAseqData")
```

To load the package

```
library(singlecellRNAseqData)
```

We now provide a brief summary of the three datasets hosted in this package and how the user can extract different features of the data from the *ExpressionSet* framework in which the data is stored.

### 2 Deng et al (2014)

---

Deng et al (2014) [?] collected single-cell RNA-seq expression data from embryos across mouse embryonic preimplantation developmental stages (zygote, 2 cell, 4 cell, 8 cell, 16 cell, early blastocyst, mid blastocyst and late blastocyst stages). The data in this package includes read counts of 22,431 genes (after removing SmartSeq2 and pooled samples) from 259 single cells. We downloaded the data from Gene Expression Omnibus (GEO:GSE45719: see <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45719>)

```
data("Deng2014MouseESC")

## Warning in data("Deng2014MouseESC"): data set 'Deng2014MouseESC' not found

Deng2014MouseESC

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 22431 features, 259 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: V278 V279 ... V205 (259 total)
##   varLabels: cell_type embryo_id
##   varMetadata: labelDescription
## featureData
```

```
## featureNames: 0610005C13Rik 0610007C21Rik ... Zzz3 (22431 total)
## fvarLabels: gene_name
## fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

Expression data for the first few genes (along rows) and the first few cells in the sample (along columns):

```
counts_deng <- Biobase::exprs(Deng2014MouseESC)
head(counts_deng[,1:5])

##           V278 V279 V280 V281 V115
## 0610005C13Rik    0    0    0    0    2
## 0610007C21Rik  194  148  378  208   26
## 0610007L01Rik 4940 5034 3714 2538  667
## 0610007P08Rik  323  672  226  241  219
## 0610007P14Rik 2501 3203 2467 1952 1195
## 0610007P22Rik   96  220  115  133   41
```

Phenotype or metadata sample information are stored in the 'phenoData' slot and include both the development stage of each cell and the corresponding embryo ID. The development stage of each cell can be extracted as follows:

```
pheno_deng <- Biobase::pData(Deng2014MouseESC)
colnames(pheno_deng)

## [1] "cell_type" "embryo_id"

table(pheno_deng$cell_type)

##
##      16cell      4cell      8cell early2cell earlyblast  late2cell  lateblast
##          50         14         28           8          43          10          30
##  mid2cell  midblast      zy
##          12         60         4
```

Gene information are stored in 'featureData' slot:

```
genes_deng <- Biobase::fData(Deng2014MouseESC)
head(genes_deng)

##           gene_name
## 0610005C13Rik 0610005C13Rik
## 0610007C21Rik 0610007C21Rik
## 0610007L01Rik 0610007L01Rik
## 0610007P08Rik 0610007P08Rik
## 0610007P14Rik 0610007P14Rik
## 0610007P22Rik 0610007P22Rik
```

### 3 Jaitin et al (2014)

---

Jaitin et al (2014) [?] collected single cell RNAseq data of mouse spleen cells using Unique Molecule Identifiers (UMIs). 4,590 single cells were included with 20,190 genes. The dataset includes in this package was downloaded from Gene Expression Omnibus (GEO:GSE54006: see <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54006>)

```
data("MouseJaitinSpleen")
MouseJaitinSpleen

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 20190 features, 4590 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: 7 8 ... 4604 (4590 total)
##   varLabels: index sequencing_batch ...
##     Column_name_in_processed_data_file (15 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 0610007C21Rik_Apr3 0610007L01Rik ... ERCC-00002 (20190
##     total)
##   fvarLabels: gene_names
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

Expression data for the first few genes (along rows) and the first few cells in the sample (along columns):

```
counts_jaitin <- Biobase::exprs(MouseJaitinSpleen)
head(counts_jaitin[,1:5])

##           7 8 9 10 11
## 0610007C21Rik_Apr3 0 0 0 1 0
## 0610007L01Rik      0 1 0 0 0
## 0610007P08Rik      0 0 0 0 0
## 0610007P14Rik      0 1 0 0 0
## 0610007P22Rik      0 0 0 0 0
## 0610009B22Rik      0 0 0 0 0
```

Phenotype or metadata of the samples include mouse ID, well ID, amplification batch, sequencing batch and can be extracted as follows:

```
pheno_jaitin <- Biobase::pData(MouseJaitinSpleen)
head(pheno_jaitin[,c("amplification_batch",
                     "sorting_markers",
                     "well_id",
```

```

                                "ERCC_dilution"]])
##      amplification_batch sorting_markers well_id ERCC_dilution
## 7                      0          CD11c+    A1      2.00E-05
## 8                      0          CD11c+    B1      2.00E-05
## 9                      0          CD11c+    C1      2.00E-05
## 10                     0          CD11c+    D1      2.00E-05
## 11                     0          CD11c+    E1      2.00E-05
## 12                     0          CD11c+    F1      2.00E-05

```

Gene names are stored in the 'featureData' slot:

```

genes_jaitin <- Biobase::featureNames(MouseJaitinSpleen)
head(genes_jaitin)
## [1] "0610007C21Rik_Apr3" "0610007L01Rik"      "0610007P08Rik"
## [4] "0610007P14Rik"      "0610007P22Rik"      "0610009B22Rik"

```

## 4 Zeisel et al (2015)

---

Zeisel et al (2015) [?] collected single cell RNA-seq data from Mouse cortex and hippocampus, with the goal of identifying different cell types. Read counts of 19,968 genes were included in the released data set for 3,005 single cells. The data in this package was downloaded from Gene Expression Omnibus (GEO:GSE60361: see <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361>)

```

data("MouseZeiselBrain")
MouseZeiselBrain
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 19968 features, 3005 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: 1772071015_C02 1772071017_G12 ... 1772058148_F03 (3005
##     total)
##   varLabels: tissue group_no ... level2_class (10 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: Tspan12 Tshz1 ... Gm20738_loc3 (19968 total)
##   fvarLabels: gene_name
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:

```

Expression data for the first few genes (along rows) and the first few cells in the sample (along columns):

```
counts_zeisel <- Biobase::exprs(MouseZeiselBrain)
head(counts_zeisel[,1:5])
```

##	1772071015_C02	1772071017_G12	1772071017_A05	1772071014_B06
## Tspan12	0	0	0	3
## Tshz1	3	1	0	2
## Fnbp1l	3	1	6	4
## Adamts15	0	0	0	0
## Cldn12	1	1	1	0
## Rxfp1	0	0	0	0

  

##	1772067065_H06
## Tspan12	0
## Tshz1	2
## Fnbp1l	1
## Adamts15	0
## Cldn12	0
## Rxfp1	0

Phenotype or metadata of the samples include tissue of origin, class type of cells, age and sex of subjects from whom the cells were extracted.

```
pheno_zeisel <- Biobase::pData(MouseZeiselBrain)
head(pheno_zeisel[,c("tissue", "sex", "age", "level1_class", "level2_class")])
```

##	tissue	sex	age	level1_class	level2_class
## 1772071015_C02	ssscortex	female	21	interneurons	Int10
## 1772071017_G12	ssscortex	male	20	interneurons	Int10
## 1772071017_A05	ssscortex	male	20	interneurons	Int6
## 1772071014_B06	ssscortex	female	21	interneurons	Int10
## 1772067065_H06	ssscortex	female	25	interneurons	Int9
## 1772071017_E02	ssscortex	male	20	interneurons	Int9

Gene names are stored in 'featureData' slot:

```
genes_zeisel <- Biobase::featureNames(MouseZeiselBrain)
head(genes_zeisel)
```

##	[1]	"Tspan12"	"Tshz1"	"Fnbp1l"	"Adamts15"	"Cldn12"	"Rxfp1"
----	-----	-----------	---------	----------	------------	----------	---------

## References

---

- [1] Jaitin DA, Kenigsberg E et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*. 343 (6172) 776-779, 2014. DOI: 10.1126/science.1247651

- [2] Deng Q, Ramskold D, Reinius B, Sandberg R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science*. 343 (6167) 193-196, 2014. DOI: 10.1126/science.1245316
- [3] Zeisel A, Munoz-Manchado AB, Codeluppi S *et al*. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*. 34: 6226, 1138-1142, 2015. DOI:10.1126/science.aaa1934